Critical events of the origins of life

Caleb Deen Bastian^{*1,2} and Herschel Rabitz^{1,3}

¹Program in Applied and Computational Mathematics, Princeton University,
Princeton, NJ. 08544, USA

²Massive Dynamics, Princeton, NJ. 08542, USA

³Department of Chemistry, Princeton University, Princeton, NJ. 08544, USA

Abstract

We discuss some critical events of the origins of life using a mathematical model and simulation studies. We find that for a replicating population of RNA molecules participating in template-directed polymerization, the hitting and establishment of a high-fidelity replicator depends critically on the polymerase fitness and sequence specificity landscapes and on genome dimension. Probability of hitting is dominated by polymerase landscape curvature, whereas hitting time is dominated by genome dimension. Surface chemistries, compartmentalization, and decay increase hitting times. These results suggest replication to be the first ‘privileged function’ marking the start of Darwinian evolution, possibly in conjunction with clay minerals or preceded by metabolism, whose dynamics evolved mostly during the final period of the search.

Index terms— Origins of life; RNA world; reaction network; stochastic simulation algorithm; random counting measure; measure-kernel-function; ordinary differential equation; stability analysis; high dimensional model representation; global sensitivity analysis; fitness and similarity sequence landscapes; hitting times; survival analysis

1 Introduction

The origins of life (OOL), abiogenesis, is a matter of high importance, for it gives insight into the distribution of life in the universe. Most OOL models are based on ‘privileged functions’ that prioritize one of replication, metabolism, or compartmentalization (Lanier and Williams, 2017). In this vein, we focus on the RNA world hypothesis, where life begins with self-replicating RNA molecules that can evolve under Darwinian evolution (Orgel, 1968; Gilbert, 1986; Joyce, 2002). This is an information-centric perspective on abiogenesis which we refer to as ‘RNA-OOL,’ representing the putative beginning of Darwinian

^{*}cbastian@princeton.edu

evolution. Other theories of abiogenesis prioritize metabolism (Dyson, 1999), cellular compartmentalization (Sakuma and Imai, 2015; Szostak et al., 2001), or hydrothermal vent chemical gradient energy (Martin et al., 2008). Information centrism interprets a living organism as an operating genetic communication system (GCS) in some connected domain that encodes and decodes genomic state relative to a replication channel.

While genomics, epigenomics, and transcriptomics of modern-day organisms are based on DNA, RNA, and epigenetic marks such as DNA methylation, RNA-OOL in its purest form concerns the dual-function of RNA as an informational polymer and ribozyme. Clues to the RNA world, among others, are found in the nucleotide moieties in acetyl coenzyme A and vitamin B12, the structure of the ribosome as a ribozyme (Cech, 2000) and moreover the centrality of RNA to the translation system, and the existence of viroids (Diener, 1971). The putative canonical RNA-OOL sequence involves RNA dependent RNA polymerase (RNAP) and Hammerhead (HH) ribozymes. It is important to note that this setup is unnecessary, as the first role of RNA may not have been template-assisted polymerization but instead based on mechanisms of RNA recombination and networking (Vaidya et al., 2012). If we assume the RNAP-HH genome is size 200 nucleotides, then there are $4^{200} \approx 10^{120}$ possible sequences in this space of genomes. Starting from some population of interacting RNA molecules, we are interested in the hitting times of critical events. Evolution seems to have concluded this search for a high-fidelity replicator in a fairly short period of time, i.e. within 400 million years of the Earth having a stable hydrosphere (Joyce, 2002).

A subtlety to the RNA-OOL argument is that template-directed polymerization by its very nature essentially requires two copies of the genome, one for the polymerase and another for the template. This makes pure RNA-OOL improbable, requiring two independent hittings. An escape hatch is the fact that the clay mineral montmorillonite, which is common on Earth, has polymerase activity for RNA (Ertem and Ferris, 1997), so minerals furnish a genome complementary to the first, enabling template-directed polymerization to proceed. Hence we assume RNA-OOL takes place in an environment having clay. Some have proposed that clay not only promotes OOL but actually constitutes it, with the first living thing clay crystals, which then later gave rise to RNA-based replication (Cairns-Smith, 1987); however, such mineral life as a GCS has high noise.

Pre-RNA worlds have been suggested, with RNA being preceded or augmented by alternative informational polymers, such as other nucleic acids (Orgel, 2000), beta amyloid (Maury, 2015), polycyclic aromatic hydrocarbons (Ehrenfreund et al., 2006), lipids (Segré et al., 2001), peptides (Kunin, 2000), and so on. Note that many of these are exposed to the so-called code conversion problem, whereby a mapping must be established from one informational alphabet to another to ensure continuity of information of the genome. RNA and DNA alphabets need only a single substitution to convert between each other (along with ribose and deoxyribose) and hence the transition from RNA to DNA in information storage is simple in principle. Furthermore it is possible that pre-RNA worlds existed independent of the RNA world in the sense they are not ancestral to the RNA world, and

that these worlds may have had non-trivial interactions with the RNA world. Thus are the possibilities.

NASA’s perspective on abiogenesis requires the following to be identified: abiotic sources of organic compounds, mechanisms of synthesis and function of macromolecules, energy sources, and environmental factors (Hays, 2015). Towards this, we assume the molecular inventories are provided by meteorites (KVENVOLDEN et al., 1970) or by Miller-Urey processes (Miller and Urey, 1959). For energy and environmental factors, we consider a variant of Darwin’s “warm little pond,” where the putative environment for origins of life is an icy pond with geothermal activity or perhaps a hydrothermal vent: ice and cold temperature facilitate complexing of single strands into double strands and polymerization (Attwater et al., 2013), and heat (energy) facilitates dissociation of double strands into single strands.

RNA-OOL has attracted many modeling efforts and analyses (Coveney et al., 2012; Lanier and Williams, 2017). Two-dimensional spatial modeling has been applied in which reactions occur locally with finite diffusion, suggesting a spatially localized stochastic transition (Wu and Higgs, 2012), simulated using Gillespie’s stochastic simulation algorithm (SSA) (Gillespie, 1977). Another model is of autocatalytic sets of collectively reproducing molecules, which has been developed in reflexively autocatalytic food-generated (RAF) theory (Hordijk and Steel, 2004). Various physics-based analyses have been conducted, such as in light of Bayesian probability, thermodynamics, and critical phenomenon (Walker, 2017). Systems of quasi-species based on the principle of natural self-organization called hypercycles employ non-equilibrium auto-catalytic reactions (Eigen and Schuster, 1977). Non-linear kinetic models for polymerization have been used to study the emergence of self-sustaining sets of RNA molecules from monomeric nucleotides (Wattis and Coveney, 1999).

We develop a simple mathematical model at the genome level to represent the synthesis and function of RNA molecules in order to gain insight into the hitting times of various critical events of the origins of life. The idea is to study the surface of hitting times in terms of the structure of the system. Let $H = \{\text{Adenine, Uracil, Guanine, Cytosine}\}$, abbreviated as $H = \{A, U, G, C\}$, corresponding to the RNA bases. We study the space of genomes having length n , $E = H^n$ with $\mathcal{E} = 2^E$, so $(E, \mathcal{E})$ is the measure space of all RNA sequences of length n with $|E| = 4^n$. Appendix A gives overview of $(E, \mathcal{E}, \nu)$. A related space, though not utilized in this article, is the space of all RNA sequences up to length n , $E^* = \cup_{i=1}^n H^i$ with $\mathcal{E}^* = 2^{E^*}$ and size $|E^*| = \frac{4}{3}(|E| - 1)$. The model lacks many features of realism, such as genome size variability, finite sources of “food” (activated nucleotides in our context), limited diffusion rates, poor system mixing, chirality, and so on, in order to concentrate on the process as a search problem. A key concept is the notion of fitness and similarity surfaces or landscapes on $(E, \mathcal{E})$. The notion of such landscapes has been studied extensively in evolutionary biology (Wright, 1984). Landscape topology has been considered in an Opti-Evo theory, which assumes sufficient environmental resources and argues that fitness landscapes do not contain “traps” and globally optimal genomes form

a connected level-set (Feng et al., 2012).

We describe the model in Section 2, where we define a replicating reaction network, whose random realizations are constructed using SSA. We describe hitting times as key random variables of interest and characterize polymerization as a transition kernel. In Section 3 we conduct and discuss simulation studies based on SSA, where we analyze the structure of the polymerase measures and the input-output and survival behaviors of the hitting times given the parameters of the system. In Section 4 we end with conclusions.

2 Materials and methods

We build a simplified mathematical model for the time-evolution of a population of interacting RNA molecules in solution. Let X_t be the population at time $t \in \mathbb{R}_+$ with initial population X_0 . X_t is a multiset. We assume the system is well mixed and has access to an infinite source of activated nucleotides.

The complement of $x \in E$ is denoted $x^c \in E$, attained using the base-pairing A with U and C with G . Let

$$h(x, y) \in \{0, 1, \dots, n\} \quad \text{for } x, y \in E \quad (1)$$

be the Hamming distance between $x, y \in E$ as the number of positions in x and y where the nucleotides differ. We have $h(x, x) = 0$ and $h(x, x^c) = n$.

2.1 Core model

We describe the reaction network of the system below.

System We model the population of genomes which can form double-stranded helices, dissociate, and noisily replicate with replicator fitness and sequence specificity. We interpret each genome as a set, either containing one element—a single-stranded genome—indicated as $\{x\}$ —or two elements, single and complementary stranded genomes, indicated as $\{x\} \cup \{x^c\} = \{x, x^c\}$ where $x, x^c \in E$. We define system elements as sets

$$\begin{aligned} \bar{E} &\equiv \{\{x\} : x \in E\} \\ \bar{F} &\equiv \{\{x, x^c\} : x \in E\} \\ \bar{G} &\equiv \bar{E} \cup \bar{F} \end{aligned}$$

with respective σ -algebras, $\bar{\mathcal{E}} = 2^{\bar{E}}$, $\bar{\mathcal{F}} = 2^{\bar{F}}$ and $\bar{\mathcal{G}} = 2^{\bar{G}}$. The reaction network of the system is given by

and expressed in terms of sets

Reaction (2) is double-strand formation from complementary single-strands. Reaction (3) is the dissociation of double-strands into single-strands, caused by a heat source. Reaction (4) is template-directed polymerization of a single-strand (the template) by another single-strand (the polymerase), producing a single-strand complementary to the template with some fidelity.

We take

$$k_{rep}(x, y) = af(x)s(x, y) \in (0, a] \quad \text{for } x, y \in E \quad (5)$$

where $a > 0$ is a positive constant, $f : E \mapsto (0, 1]$ is the replicative *fitness* of x (as a polymerase) and $s : E \times E \mapsto (0, 1]$ is the *similarity* between x and y , a symmetric function. The idea is that the replicator has fitness and specificity for sequences. Finally x replicates y , outputting a noisy version y_*^c , where each nucleotide position has fidelity probability $p : E \mapsto (0, 1]$. This is a continuous-time branching model with mutation. We simulate trajectories of the system using SSA (Gillespie, 1977).

The system is assumed to be well-mixed, and all parts of the system are accessible for reactions. This means that reactions across various disjoint volume elements of the system are dependent. We do not consider the effects of finite diffusion, which effects a length scale above which disjoint volume elements are effectively independent (Wu and Higgs, 2012). Basically limited diffusion breaks the system into non-interacting regions, whereby reactions are localized.

High-fidelity set Pick an arbitrary subset of genomes $R \subset E$ as high-fidelity replicators of size $r = |R|$. We define R two ways.

We define R using a product of non-empty random nucleotide subsets $\{A_i \subseteq H : i = 1, \dots, n\}$ for each nucleotide position

$$R = A_1 \times \dots \times A_n$$

so that

$$r = \prod_{i=1}^n |A_i| = 2^{r_2} 3^{r_3} 4^{r_4}$$

where $r_2 = |\{A_i : |A_i| = 2\}|$, etc. and $r_1 + r_2 + r_3 + r_4 = n$. For simplicity, we assume $|A_i| \in \{1, 4\}$ with fraction 4 being $q \in (0, 1)$. Thus R is a subset of E defined as a product space. This model is essentially single-nucleotide polymorphism (SNP), where each SNP position takes values of any of the nucleotides.

For another construction of R , we define a finite union of m random genomes $R = \{x_1, \dots, x_m\}$.

Distance Define the Hamming distance H between x and R as

$$H(x, R) = \min\{h(x, y) : y \in R\} \in \{0, 1, \dots, n\} \quad \text{for } x \in E. \quad (6)$$

Fitness Define “tent-pole” fitness f_k of $x \in E$ and R for $k \in \mathbb{R}_+$ as

$$f_k(x, R) = \exp[-k H(x, R)] \in (0, 1] \quad \text{for } x \in E. \quad (7)$$

The maximums are the genomes of R , which are the “points” or “poles” of the surface, with exponential decay into the remainder of the space in string distance. The strength of the decay is governed by parameter k , which can be specified through the value of fitness at $H(x, R) = n$ (genome dimension),

$$k = -\frac{\log(f_k(x, R))}{n} \quad (8)$$

Appendix B describes other fitness functions.

Similarity Symmetric distance S is defined as

$$S(x, y) = h(x, y) \wedge h(x, y^c) \wedge h(x^c, y) \wedge h(x^c, y^c) \in \{0, 1, \dots, n\} \quad \text{for } x, y \in E. \quad (9)$$

Similarity s_k of $x, y \in E$ for $k \in \mathbb{R}_+$ is defined in terms of exponential decay as

$$s_k(x, y) = \exp[-k S(x, y)] \in (0, 1] \quad \text{for } x, y \in E. \quad (10)$$

Similarity is a measure of the distance between the replicator and the template or template complement, where smaller distances convey higher similarities. Various parameterizations of similarity may be considered.

Fidelity Finally, polymerization fidelity probability for $k \in \mathbb{R}_+$ is defined as

$$p_k(x, R) = f_k(x, R) \in (0, 1] \quad \text{for } x \in E. \quad (11)$$

Note that $f_k(x, R) = p_k(x, R) = 1$ for $x \in R$.

Note that fitness, similarity, and fidelity are defined for single-stranded genomes $(E, \mathcal{E})$.

Counting representation The process $X = (X_t)_{t \in \mathbb{R}_+}$ is the time-evolution of the system. X_t contains the individual single stranded molecules in the set $\bar{E}$ and double stranded molecules in the set $\bar{F}$, with set $\bar{G} = \bar{E} \cup \bar{F}$ having size $m = |\bar{G}| = 3|\bar{E}|/2$. Note that $\{x, x^c\} = \{x^c, x\}$, so $|\bar{F}| = |\bar{E}|/2$. X_t induces a random counting measure N_t on $(\bar{G}, \bar{\mathcal{G}})$ as

$$N_t(A) = \sum_{x \in X_t} \mathbb{I}_A(x) \quad \text{for } A \in \bar{\mathcal{G}} \quad (12)$$

The total count is $K_t \equiv |X_t| = N_t(\bar{G})$. We assume that the counter N_t is maintained for all $t \in \mathbb{R}_+$.

Reaction rates The total reaction rate is given by

$$\check{k}(t) = \check{k}_{ds}(t) + \check{k}_{ss}(t) + \check{k}_{rep}(t) \quad \text{for } t \in \mathbb{R}_+$$

where

$$\check{k}_{ds}(t) = \frac{1}{2} \sum_{\{x\} \in \bar{E}} k_{ds} N_t(\{x\}) N_t(\{x^c\})$$

and

$$\check{k}_{ss}(t) = \sum_{\{x, x^c\} \in \bar{F}} k_{ss} N_t(\{x, x^c\})$$

and

$$\check{k}_{rep}(t) = \sum_{(\{x\}, \{y\}) \in \bar{E}^2} k_{rep}(x, y) N_t(\{x\}) (N_t(\{y\}) - \mathbb{I}(x = y))$$

The first reaction rate is a sum over $\bar{E}$ with size 4^n . The second a sum over $\bar{F}$ with size $4^n/2$. The third reaction is a sum over $\bar{E}^2$ with $4^{2n} = 16^n$ elements. Therefore $\check{k}(t)$ requires $4^n(\frac{3}{2} + 4^n)$ elements to be evaluated for every reaction. Clearly direct representation on the full space is very expensive and impractical for even modest n . One obvious way to improve efficiency is not summing over the zero elements. Form sets

$$\mathcal{X}_1 = \{x \in \text{set}(X_t) : |x| = 1\} \quad (13)$$

and

$$\mathcal{X}_2 = \{x \in \text{set}(X_t) : |x| = 2\} \quad (14)$$

Then direct calculations are

$$\check{k}_{ds}(t) = \frac{1}{2} \sum_{\{x\} \in \mathcal{X}_1} k_{ds} N_t(\{x\}) N_t(\{x^c\})$$

and

$$\check{k}_{ss}(t) = \sum_{\{x, x^c\} \in \mathcal{X}_2} k_{ss} N_t(\{x, x^c\})$$

and

$$\check{k}_{rep}(t) = \sum_{(\{x\}, \{y\}) \in \mathcal{X}_1^2} k_{rep}(x, y) N_t(\{x\})(N_t(\{y\}) - \mathbb{I}(x = y)).$$

For this approach the replication rate has quadratic dependence on $|\mathcal{X}_1|$. Using the reaction rates, the system may be exactly simulated using SSA. The reaction at time t with rate $\check{k}(t)$ occurs over time interval $\Delta t \sim \text{Exponential}(1/\check{k}(t))$. As $\check{k}(t)$ increases with increasing $K_t = N_t(\bar{G})$, the reaction rate increases and reaction duration Δt decreases over time. The natural consequence of increasing process intensity is that the system speeds up.

The quadratic dependence may still be too expensive for large simulations. Appendix D describes Monte Carlo approximation of the reaction rates.

2.2 Hitting times

We define some hitting times. The initial population consists of I single-stranded genomes X_0 . Define hitting time τ for the time of the first replication event

$$\tau_{rep} = \inf \{t \in \mathbb{R}_+ : K_t > I\}. \quad (15)$$

Define the hitting time τ for the appearance of R ,

$$\tau_R = \inf \{t \in \mathbb{R}_+ : \mathbb{I}_{\{1,2,\dots\}}(|R \cap X_t|) = 1\}. \quad (16)$$

Put $X_R = R \cup \{x, x^c : x \in R\}$ and define the volume fraction of R at time t as

$$V(t) = \frac{N_t(X_R)}{K_t}. \quad (17)$$

Define hitting time τ where R emerges and reaches a minimum volume fraction,

$$\tau_{min} = \inf \{t \in \mathbb{R}_+ : t \geq \tau_R, V(t) \leq V(s) \text{ for } \tau_R \leq s \leq t\}. \quad (18)$$

Define hitting time τ for the time R constitutes some volume fraction $v \in (0, 1]$ of the population,

$$\tau_v = \inf \{t \in \mathbb{R}_+ : V(t) \geq v\}. \quad (19)$$

In practice for simulations, τ_v is censored based on some total number of reactions, that is, if the volume fraction is not achieved by n reactions, $\tau_v = \infty$ because there is no arrival time.

We have that $\tau_{rep} \leq \tau_R \leq \tau_{min} \leq \tau_v$.

For SSA, we specify a maximum number of reactions N to simulate. We have parameters $\theta \in \Theta$ for τ , such as landscape curvature k , genome dimension n , etc. Therefore $\tau(\theta)$ is right-censored with value ∞ at simulation time a , as some simulations will stop at time a with no arrival time. These are censoring events. For fixed θ the $\tau(\theta)$ is a random variable, due to the stochastic nature of SSA. Hence for each θ , we attain a set of M realizations of τ as

$$\mathcal{T}(\theta) = \{\tau_i(\theta) : i = 1, \dots, M\}. \quad (20)$$

For convenience we assume the $\mathcal{T}(\theta)$ is ordered by non-censored followed by censored.

Functional structure For each θ , we record two values: the number of hitting events in $\mathcal{T}$

$$g(\theta) = |\{x \in \mathcal{T}(\theta) : x < \infty\}| \in \{0, 1, \dots, M\} \quad (21)$$

and the average of τ

$$f(\theta) = \begin{cases} \frac{1}{g(\theta)} \sum_{i=1}^{g(\theta)} \tau_i(\theta) & \text{if } g(\theta) > 0 \\ 0 & \text{if } g(\theta) = 0 \end{cases} \in \mathbb{R}_+ \quad (22)$$

To describe the functional structure of $f(\theta)$, we require a classifier which determines whether or not $g(\theta) = 0$ and a regressor for the value of $f(\theta)$ for $g(\theta) > 0$. Assume that the parameters $\theta = (\theta_1, \dots, \theta_n)$ are randomly sampled according to distribution $\nu = \prod_i \nu_i$ and the hitting times recorded. High dimensional model representation (HDMR) may be attained for the classifier (as a probabilistic discriminative model) and the regressor of $f(\theta)$. For the regressor, we have HDMR

$$f(\theta_1, \dots, \theta_n) = f_0 + \sum_i f_i(\theta_i) + \sum_{i < j} f_{ij}(\theta_i, \theta_j) + \dots + f_{1\dots n}(\theta_1, \dots, \theta_n)$$

The HDMR component functions $\{f_u\}$ convey a global sensitivity analysis, where defining variance term

$$\sigma_u^2 = \text{Var} f_u = \int_{\Theta} f_u^2(\theta_u) \nu(d\theta)$$

we have a decomposition of variance

$$\sigma_f^2 = \text{Var} f = \sum_{u \subseteq \{1, \dots, n\} : |u| > 0} \sigma_u^2$$

The normalized terms $S_u = \sigma_u^2 / \sigma_f^2$ are called sensitivity indices. Appendix E gives brief description of global sensitivity analysis via HDMR.

Statistical structure A second analysis can be conducted on $\mathcal{T}(\theta)$ for the θ using reliability theory. Put $\mathfrak{T}(\Theta) \equiv \{\mathcal{T}(\theta) : \theta \in \Theta\}$ where $\Theta = \{\theta_i\}$ is an independency of parameter values. For each $\theta \in \Theta$ we partition $\mathcal{T}(\theta)$ into C censored values with censor times $\mathcal{C}(\theta) = \{a_i(\theta)\}$ and $M - C$ non-censored (hitting) values $\mathcal{N}(\theta) = \{x \in \mathcal{T}(\theta) : x < \infty\}$. The likelihood is given by

$$L(\mathfrak{T}(\Theta)|\vartheta) = \prod_{\theta \in \Theta} \prod_{x \in \mathcal{N}(\theta)} f(x|\vartheta) \prod_{x \in \mathcal{C}(\theta)} R(x|\vartheta)$$

where f is the hitting time probability density function ('failure density') and R is censoring time distribution ('reliability distribution') and ϑ are the parameters of the density and

distribution functions. Note that f and R each specify each other, so ϑ are the common parameters. Reliability definitions are given in Appendix F.

We show reliability quantities in Table 1 for the two-parameter $(\alpha, \beta) \in (0, \infty)^2$ Weibull(α, β) distribution and Cox proportional hazard's model where γ is a vector of coefficients for the θ . We use the Python software *lifelines* for estimation of ϑ for the Weibull-Cox model from data (Davidson-Pilon, 2019). The mean failure time $\mathbb{E}\tau(\theta)$ is given by

$$\mathbb{E}\tau(\theta) = \int_0^\infty t f(t, \theta | \alpha, \beta, \gamma) = \beta e^{-\gamma \cdot \theta / \alpha} \Gamma(1 + \frac{1}{\alpha})$$

We have second moment

$$\int_0^\infty t^2 f(t, \theta | \alpha, \beta, \gamma) = \beta^2 e^{-2\gamma \cdot \theta / \alpha} \Gamma(1 + \frac{2}{\alpha})$$

giving variance

$$\text{Var}\tau(\theta) = \beta^2 e^{-2\gamma \cdot \theta / \alpha} (\Gamma(1 + \frac{2}{\alpha}) - \Gamma^2(1 + \frac{1}{\alpha}))$$

Thus if $\gamma < 0$, then $\mathbb{E}\tau(\theta)$ and $\text{Var}\tau(\theta)$ exponentially increase in θ .

[Table 1]

2.3 Hitting cardinality

We index the N reactions of X_t with arrival times $\{T_i\}$ in $(\bar{\mathbb{R}}_+, \mathcal{B}_{\bar{\mathbb{R}}_+})$ where $\bar{\mathbb{R}} = \mathbb{R}_+ \cup \{\infty\}$. Define the hitting reaction $\varpi(\theta) \in \mathbb{N}$ in terms of hitting time $\tau(\theta) \in \bar{\mathbb{R}}_+$ as

$$\varpi(\theta) = \sum_{i=1}^N \mathbb{I}_{[0, \tau(\theta)]}(T_i)$$

$\varpi(\theta)$ is right-censored at N reactions. If $\tau(\theta) = \infty$, then $\varpi(\theta) = N$. If $\tau(\theta) < \infty$ then $\varpi(\theta) < N$.

2.4 Surface chemistries

The system given by (2) of noisy polymerization requires two separate hitting events, one $x \in R$ and either $x \in R$ or $x^c \in E$ in order for high-fidelity replicators to maximally engage in templated-directed polymerization and achieve some fraction of the population. This setup of RNA polymerase action, requiring two such events for the polymerase and template, makes the hitting times long. Basically the same information must be discovered twice before it can be used, which is unsatisfactory. Other molecules have RNA oligomerization and polymerase activities, such as the clay montmorillonite (Ferris, 2006; Ertem and Ferris, 1997), that are naturally found on Earth in various environments. Here the

surface chemistry of montmorillonite performs oligomerization and polymerization of RNA molecules. The reactions are given by

where polymerization is noisy with fidelity probability $p \in (0, 1]$ and x_*^c is the noisy complement of x . The reaction rates are given by

$$\check{k}_{clay-o}(t) = k_{clay-o}$$

and

$$\check{k}_{clay-p}(t) = k_{clay-p} N_t(\mathcal{X}_1)$$

Therefore, upon the first hitting of R with genome $x \in R$ through (4) or (23), x gives two high-similarity single-stranded genomes x and x_*^c through (24), which then may participate in template-directed RNA polymerization (4).

2.5 Reactions as measure-kernel-functions

All the reactions $x \mapsto y$ which involve substrate x may be represented using transition kernels, which form linear operators. At each iteration of SSA, a reaction type is chosen, followed by a transition to a particular domain $(X, \mathcal{X})$ with distribution ν_t , followed by mapping into a codomain $(Y, \mathcal{Y})$ using transition probability kernel Q with distribution $\mu_t = \nu_t Q$. The notions of ν_t and Q involve *measure-kernel-functions*. The probability of transition of y into $B \in \mathcal{Y}$ given x is given by Q

$$Q(x, B) = \int_B Q(x, dy) = \mathbb{P}(y \in B : x)$$

Appendix C recalls some facts about Q .

We index the reaction types on $(Z, \mathcal{Z}) = (\mathbb{N}_{>0}, 2^{\mathbb{N}_{>0}})$. Let η_t be the probability measure on $(Z, \mathcal{Z})$ formed from the normalized reaction rates. Let Q_* be the transition kernel from $(Z, \mathcal{Z})$ into $(X, \mathcal{X})$. Then $\eta_t Q_* = \nu_t$ is the distribution on $(X, \mathcal{X})$ and $\mu_t = \nu_t Q$ is the distribution on $(Y, \mathcal{Y})$. Then a reaction is the mapping $(Z, \mathcal{Z}) \mapsto (X, \mathcal{X}) \mapsto (Y, \mathcal{Y})$.

We define kernels Q for RNA and clay polymerization to provide insight into the reactions. Consider X_t for some $t \in \mathbb{R}_+$. Recall N_t is the random counting measure of X_t on single and double-stranded genomes $(E \cup F, 2^{E \cup F})$. For RNA and clay polymerization we take ν_t as a (random) probability measure for x in domain $(X, \mathcal{X})$ and describe a transition probability kernel Q from x into y in codomain $(Y, \mathcal{Y})$.

RNA polymerization For RNA polymerization we have that

$$\{x\}, \{y\} \mapsto \{x\}, \{y\}, \{y_*^c\}$$

which results in the creation of the single-stranded genome $\{y_*^c\}$. The first dimension is the polymerase and the second is the template.

Definition 1 (Measure on domain ν_t). *Let ν_t be a random probability measure on $(E \times E, \mathcal{E} \otimes \mathcal{E})$ formed by random counting measure N_t (12)*

$$\nu_t\{x, y\} = \frac{k_{rep}(x, y)N_t(\{x\})(N_t(\{y\}) - \mathbb{I}(x = y))}{\check{k}_{rep}(t)} \quad \text{for } (x, y) \in E \times E$$

with

$$\nu_t f = \sum_{(x, y) \in E \times E} \nu_t\{x, y\} f \circ (x, y) \quad \text{for } f \in (\mathcal{E} \otimes \mathcal{E})_+ \quad (25)$$

We write $\nu_t(A) = \nu_t \mathbb{I}_A$ for $A \in \mathcal{E} \otimes \mathcal{E}$.

Because the first two coordinates are preserved under the mapping, we focus on the new dimension as a transition from $(E \times E, \mathcal{E} \otimes \mathcal{E})$ into $(E, \mathcal{E})$ using transition probability kernel Q . In this case, Q is defined by a $16^n \times 4^n$ matrix whose rows vectors (dimension 4^n) are probability vectors. The structure of Q follows from the polymerase noisy replication, whereby each nucleotide position has fidelity probability $p : E \mapsto (0, 1]$, which depends on the first dimension of $E \times E$. We put $p_x = p(x)$ for $x \in E$.

Theorem 1 (Error distribution). *The number of errors by polymerase $x \in E$ on template $y \in E$ is distributed*

$$h(y^c, y_*^c) \sim \text{Binomial}(n, 1 - p_x) \quad \text{for } p_x \in (0, 1)$$

with mean $n(1 - p_x)$ and variance $np_x(1 - p_x)$ and

$$h(y^c, y_*^c) \sim \text{Dirac}(0) \quad \text{for } p_x = 1.$$

Partition E into level sets $(H_i(y))$ by Hamming distance to the template complement y^c ,

$$H_i(y) = \{x \in E : h(y^c, x) = i\} \quad \text{for } i \in \{0, \dots, n\}. \quad (26)$$

We define Q for RNA polymerization. Q completely encodes RNA polymerization using Theorem 1.

Corollary 1 (Transition probability kernel Q). *We have that*

$$Q \quad (27)$$

for RNA polymerization is defined by

$$Q((x, y), H_i(y)) = \binom{n}{i} (1 - p_x)^i p_x^{n-i} \quad \text{for } (x, y) \in E \times E, \quad i \in \{0, \dots, n\}, \quad p_x \in (0, 1)$$

and

$$Q((x, y), \{z\}) = \frac{1}{|H_i(y)|} \binom{n}{i} (1 - p_x)^i p_x^{n-i} \quad \text{for } (x, y) \in E \times E, \quad i \in \{0, \dots, n\}, \quad z \in H_i(y), \quad p_x \in (0, 1)$$

and

$$Q((x, y), \{y^c\}) = 1 \quad \text{for } (x, y) \in E \times E, \quad p_x = 1.$$

RNA polymerization is defined by Q using the binomial structure of polymerase error. A more sophisticated model could be defined as a sum of Bernoulli random variables with varying success probabilities in the Poisson binomial distribution. This could be used to take into account polymerase error that varies with nucleotide position, such as taking into account schemata such as repeats which destabilize the polymerase (Bornholt et al., 2016).

Proposition 1 (Measure on codomain μ_t). $\mu_t = \nu_t Q$ is a probability measure on $(E, \mathcal{E})$ defined by

$$\mu_t f = \int_{E \times E} \nu_t(dx, dy) \int_E Q((x, y), dz) f(z) \quad \text{for } f \in \mathcal{E}_+ \quad (28)$$

It is multiplication of ν_t as a 16^n dimension row vector with $16^n \times 4^n$ dimension matrix Q , giving a 4^n dimension row vector $\nu_t Q$. We write $\mu_t(A) = \mu_t \mathbb{I}_A$ for $A \in \mathcal{E}$.

Define the partition (H_i) of E as

$$H_i \equiv \{x \in E : \min\{H(x, R), H(x^c, R)\} = i\} \quad \text{for } i \in \{0, \dots, n\}. \quad (29)$$

Then $\mu_t(H_i)$ for $i \in \{0, \dots, n\}$ is the distribution on genomes by distance to R , i.e.

$$\mu_t(H_i) = \sum_{x \in E} \mu_t\{x\} \mathbb{I}_{H_i}(x) \quad \text{for } i \in \{0, \dots, n\}$$

contains the instantaneous information of RNA polymerization.

Clay polymerization For clay polymerization, we have that

$$\{x\} \mapsto \{x\}, \{x_*^c\}$$

which we regard as a mapping from $(E, \mathcal{E})$ into $(E, \mathcal{E})$. Let ν_t be a probability measure on $(E, \mathcal{E})$ defined by

$$\nu_t\{x\} = \frac{N_t(\{x\})}{N_t(E)} \quad \text{for } x \in E$$

Similar to RNA polymerization, for clay fidelity probability $p \in (0, 1]$ we have Q as the $4^n \times 4^n$ matrix defined by

$$Q(x, H_i(x)) = \binom{n}{i} (1-p)^i p^{n-i} \quad \text{for } x \in E, \quad i \in \{0, \dots, n\}, \quad p \in (0, 1)$$

and

$$Q(x, \{z\}) = \frac{1}{|H_i(x)|} \binom{n}{i} (1-p)^i p^{n-i} \quad \text{for } x \in E, \quad i \in \{0, \dots, n\}, \quad z \in H_i(x), \quad p \in (0, 1)$$

and

$$Q(x, \{x^c\}) = 1 \quad \text{for } x \in E, \quad p = 1.$$

$\mu_t = \nu_t Q$ is a probability measure on $(E, \mathcal{E})$ defined by

$$\mu_t f = \int_E \nu_t(dx) \int_E Q(x, dy) f(y) \quad \text{for } f \in \mathcal{E}_+$$

Note that, for SSA, Q is fixed over the simulation, whereas ν_t depends on time. That is, the reactions are chosen according to the reaction rates, and the reactions each use respective Q . The ν_t is formed using a random counting measure, so ν_t is random. This approach generalizes in the obvious way to all the reactions.

2.6 Deterministic model

Consider the core model defined in (2), (3), and (4) with reaction rates $\check{k}_{ds}(t)$, $\check{k}_{ss}(t)$, and $\check{k}_{rep}(t)$. We are neglecting clay and decay for the moment. All the reactions impact $(E, \mathcal{E})$. For double-strand formation, the input space is $(X, \mathcal{X}) = (E \times E, \mathcal{E} \otimes \mathcal{E})$ and the output space is $(Y, \mathcal{Y}) = (F, \mathcal{F})$. For double-strand dissociation, the input and output spaces are swapped. For polymerization, $(X, \mathcal{X}) = (E \times, \mathcal{E} \otimes \mathcal{E})$ and $(Y, \mathcal{Y}) = (E, \mathcal{E})$. Hence $(E, \mathcal{E})$ is positively impacted by polymerization, positively impacted by double-strand dissociation, and negatively impacted by double-strand formation. Put $[x] = N_t(\{\{x\}\})$ and $[x, x^c] = N_t(\{\{x, x^c\}\})$. Recall the ν_t , Q , and $\mu_t = \nu_t Q$ for the reactions, e.g., ν_t^{rep} , ν_t^{ds} , Q^{rep} , etc.

For $x \in E$, we have the system of $m = \frac{3}{2}4^n$ deterministic non-linear ordinary differential equations (ODEs) in m variables as mean-field equations

$$\begin{aligned} \frac{d[x]}{dt} &= f_x = \check{k}_{rep}(t)\mu_t^{rep}\{\{x\}\} + \check{k}_{ss}(t)\mu_t^{ss}\{\{x\}, \{x^c\}\} - \check{k}_{ds}(t)\mu_t^{ds}\{\{x, x^c\}\} \\ &= \check{k}_{rep}(t) \sum_{(y,z) \in E \times E} \nu_t^{rep}\{\{y\}, \{z\}\} Q^{rep}((y, z), \{\{x\}\}) + k_{ss}[x, x^c] - k_{ds}[x][x^c] \\ &= \sum_{(y,z) \in E \times E} k_{rep}(y, z)[y]([z] - \mathbb{I}(y = z))Q^{rep}((y, z), \{\{x\}\}) + k_{ss}[x, x^c] - k_{ds}[x][x^c] \\ \frac{d[x, x^c]}{dt} &= f_{xx^c} = k_{ds}[x][x^c] - k_{ss}[x, x^c] \end{aligned} \quad (30)$$

The fixed points of f are the equilibria of the system, i.e. $f(x) = \mathbf{0}$ for $x \in \mathbb{R}_+^m$. The Jacobian of the system is

$$J = \begin{bmatrix} \frac{df_1}{dx_1} & \cdots & \frac{df_1}{dx_m} \\ \vdots & \ddots & \vdots \\ \frac{df_m}{dx_1} & \cdots & \frac{df_m}{dx_m} \end{bmatrix} \quad (31)$$

The eigenvalues of the Jacobian reveal the stability of the fixed points. If all the eigenvalues of the Jacobian evaluated at the fixed point have negative real part, then the fixed point is stable. If none of the eigenvalues are zero and at least one of the eigenvalues has positive real part, then the fixed point is unstable. If at least one eigenvalue is zero, then the fixed point can be either stable or unstable.

2.7 Decay

The RNA genomes have finite lifetimes in reality. This comes from a variety of sources, including radiation, pH, intrinsic molecular stability, etc. Assume double-stranded RNA is stable, whereas single-stranded RNA is not. Therefore we create a reaction for decay of single-stranded RNA into constitutive nucleotides

with reaction rate

$$\check{k}_\emptyset(t) = k_\emptyset N_t(\mathcal{X}_1)$$

2.8 Compartmentalization

It is thought that compartmentalization plays a role in the origins of life, giving foci of reproducing genomes (Kitadai and Maruyama, 2018; Szostak et al., 2001). This is somewhat anticipated by the similarity measure s , where genomes are more likely to copy

similar genomes than less similar ones. Explicit spatial effects may be modeled by assuming each $x \in E$ is marked with a position on a bounded subset of the real line $([-T, T], \mathcal{B}_{[-T, T]}) \subset (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. We think of this as a one-dimensional projection of the three-dimensional system. Additional species can be introduced, such as lipids, with reactions forming a micelle M (micellisation), which encloses some $A = [r, s] \subset [-T, T]$. We assume the lipids interact with the single stranded genomes in A to form micelles as

with reaction rate

$$k_{mic}(t) = k_{mic} N_t(\mathcal{X}_1).$$

Hence micellisation is coupled to the population of genomes by design so that it evolves on roughly the same time-scale as genome activities. Note that micelles can enclose one another, i.e. $M(A)$ and $M(B)$ where $A \subset B$ or $B \subset A$, but cannot cross, i.e. for all micelles at locations $A, \dots, B$ we have that $A \cap B \in \{A, B, \emptyset\}$. For example, suppose one micelle $A = [0, 1]$ encloses another two, $B = [\frac{1}{3}, \frac{1}{2}]$ and $C = [\frac{2}{3}, \frac{3}{4}]$. Then $A \setminus (B \cup C) = [0, \frac{1}{3}) \cup (\frac{1}{2}, \frac{2}{3}) \cup (\frac{3}{4}, 1]$. Although $A \setminus (B \cup C)$ is disconnected in one dimension, the intervals are physically connected in three dimensions, where micelles are spheres. We identify each micelle to a union of disjoint intervals, disjoint across the micelles.

Let $\mathfrak{M} = \{A \subset [-T, T]\}$ be the collection of disjoint micelle regions. Let $l : G \mapsto [-T, T]$ be the location function of the genome. Let $L_t = \{l(x) : x \in X_t\}$ be the location set of the genomes. Then the random measure N_t on $(G, \mathcal{G})$ is expanded to M_t on $(G \times [-T, T], \mathcal{G} \otimes \mathcal{B}_{[-T, T]})$,

$$M_t(A) = \sum_{(x, l) \in (X_t, L_t)} \mathbb{I}_A(x, l) \quad \text{for } A \in \mathcal{G} \otimes \mathcal{B}_{[-T, T]}$$

2.9 Metabolism

Metabolism is often promoted in origins-of-life (Dyson, 1999). For our purposes here, we shall identify metabolism reaction-state to the measurable space $(M, \mathcal{M})$ with probability measure ν . Let Q_m be a transition kernel from $(M, \mathcal{M})$ into $(E, \mathcal{E})$, positing that metabolism precedes replication. For example, certain metabolic state may be precursors to the synthesis of RNA. Consider product space $(M \times E, \mathcal{M} \otimes \mathcal{E})$ with measure $\mu = \nu \times Q_m$. Now we suppose that, upon achieving Darwinian evolution in replicators, the replicators will eventually become adapted to $(M, \mathcal{M})$. Hence we interpret $(M, \mathcal{M})$ as a mark-space of $(M \times E, \mathcal{M} \otimes \mathcal{E})$, representing genomic adaptation. Let Q_e be a transition kernel from $(M \times E, \mathcal{M} \otimes \mathcal{E})$ into $(M, \mathcal{M})$. Then $\mu = \nu \times Q_m \times Q_e$ is a probability measure on $(M \times E \times M, \mathcal{M} \otimes \mathcal{E} \otimes \mathcal{M})$, where

$$\mu f = \int_M \nu(dx) \int_E Q_m(x, dy) \int_M Q_e((x, y), dz) f(x, y, z) \quad \text{for } f \in (\mathcal{M} \otimes \mathcal{E} \otimes \mathcal{M})_+$$

or

$$\mu(dx, dy, dz) = \nu(dx)Q_m(x, dy)Q_e((x, y), dz).$$

Therefore metabolism-first followed by replication and genomic adaptation is encoded by the structures of Q_m and Q_e . We do not specify these transition kernels in this article but comment they are richly textured.

2.10 Reaction overview

The reactions of the system having decay and clay and their reaction orders are shown in Table 2. There is one zero-order reaction, three first-order reactions, and two second-order reactions. Additionally we show reactions and orders for compartmentalization and metabolism.

[Table 2]

3 Results

Consider some initial population of I random genomes X_0 . The population over time is given by X_t with associated random counting measure N_t on $(\bar{G}, 2^{\bar{G}})$. Recall parameters

$$\theta = (n, q, k, l, m, p, k_{\emptyset}, k_{ss}, k_{ds}, k_{rep}, k_{clay-o}, k_{clay-p})$$

for genome dimension n , high-fidelity genome set size q , fitness degree k , similarity degree l , fidelity degree m , clay fidelity probability p , RNA decay rate $k_{\emptyset}$, double-strand dissociation rate k_{ss} , double-strand formation rate k_{ds} , RNA replication rate k_{rep} , and clay oligomerization and polymerization rates k_{clay-o} and k_{clay-p} . These parameters are summarized in Table 3.

[Table 3]

With the parameters governing the reaction rates, different values of these parameters confer different regimes for the system.

3.1 Stability: ODEs

We characterize the zeros of the vector field f from ODE system (30) and use the eigenvalues of the Jacobian (31) to determine their stability.

Theorem 2. *The ODE system (30) for $R = \{x\}$, $x \in E$, has a single unstable fixed-point at $[x] = 1$ and $[y] = 0$ for $y \in G \setminus x$.*

Proof. Solving $f = 0$ gives a single solution $[x] = 1$ and $[y] = 0$ for $y \in G \setminus x$. For this solution, the eigenvalues of the Jacobian contain no zero values and positive values. Therefore the solution is unstable. $\square$

It follows from Theorem 2 that for all other initial conditions, the system has no equilibria.

Corollary 2. *For all initial conditions X_0 such that $I = |X_0| > 1$, the system is unbounded.*

This confirms the obvious: the system, a replicating network with no death, is almost always an increasing system.

3.2 Simulation reaction state

We are interested in the behavior of ν_t (25) and $\mu_t = \nu_t Q$ (28) for RNA polymerization. These reveal the instantaneous information of the system. The structure of μ_t reveals the state of polymerization and is a leading indicator of the population concentrations over time.

3.2.1 Core model with “tent” functions, probable hitting $\mathbb{P}(\tau_v(\theta) < \infty) \sim 1$

Take genome dimension $n = 3$ and fitness and similarity parameters $k = l = -\log(0.01)/n$ and fidelity parameter $m = -\log(0.25)/n$. Set rates $k_{ss} = k_{ds} = 1$ and $k_{rep} = 10$ and use the “tent” function for fitness, similarity, and fidelity. Take random X_0 with $I = |X_0| = 10$ and random singleton $R = \{\{x\}\}$ ($q = 0$). We simulate 5 000 reactions, simulation censored at τ_v for $v = 0.25$. Take partition (H_i) (29) of E . In Figure 1 we plot measures of a typical realization of X_t on (H_i) of concentration (Figure 1a), growth (Figure 1b), and μ_t (Figure 1c). Some quantities are plotted on log-log scale, whereas others are plotted on a linear-log scale. These results show that the concentrations are relatively stable for most time, until the high-fidelity manifold is hit. Then the concentration of high-fidelity replicators rapidly increases to exceed 25%. Similarly, Figure 1b shows the growth curves on a log-log scale, where the high-fidelity manifold rapidly increases near the end of the simulation. Figure 1c shows the structure of μ_t . Low probability is assigned to polymerization of high-fidelity replicators for most of the reaction time, followed by a large increase near the end of the simulation, where high-fidelity replicators dominate with 56% probability. *Therefore μ_t is a leading indicator of the concentration curve, i.e. at simulation end-time, concentration of high-fidelity replicators is 25% and polymerization output is 56%.*

[Figure 1]

3.2.2 Core model with “tent” functions, improbable hitting $\mathbb{P}(\tau_v(\theta) < \infty) \sim 0$

We use the same configuration as Section 3.2.1 except for setting $k = l = -\log(0.1)/n$. In Figure 2 we plot measures of a typical realization of X_t on (H_i) of concentration (Figure 2a), growth (Figure 2b), and μ_t (Figure 2c). The behavior has completely changed: the high-fidelity group ends the simulation with around 6% concentration, only steadily increasing,

and never hits. The polymerase output μ_t shows 6%. This indicates that the concentration of high-fidelity replicators is unlikely to increase further, as the population is generally in equilibrium with the polymerase output.

[Figure 2]

3.2.3 Core model with linear functions, improbable hitting $\mathbb{P}(\tau_v(\theta) < \infty) \sim 0$

The same configurations for Section 3.2.1 are used, except the fitness, similarity, and fidelity functions are linear. Similar to the “tent” functions, we specify the terminus k . Then

$$f_k(x, R) = 1 + \left(\frac{k-1}{n} \right) H(x, R) \quad \text{for } x \in E$$

and

$$s_k(x, y) = 1 + \left(\frac{k-1}{n} \right) S(x, y) \quad \text{for } x, y \in E$$

We put $k = 0.01$ for fitness and similarity and $k = 0.25$ for fidelity. We simulate X_t for 5 000 reactions. We find that $\mathbb{P}(\tau_{0.25}(\theta) < \infty) \sim 0$. In Figure 3 we plot measures of a typical realization of X_t on (H_i) of concentration (Figure 3a), growth (Figure 3b), and μ_t (Figure 3c). The simulation ends with high-fidelity concentration of $\sim 5\%$ and polymerase output of $\sim 4\%$. Therefore the concentration of high-fidelity replicators will continue to decrease. *Linear surfaces are not sufficient to achieve hitting times $\tau_v(\theta) < \infty$ for $v = 0.25$, in contrast to the non-linear “tent” functions.*

[Figure 3]

3.2.4 Expanded model with “tent” functions, probable hitting $\mathbb{P}(\tau_v(\theta) < \infty) \sim 1$

We consider similar model to the previous subsections and expand it with clay oligomerization rate (of RNA) k_{clay-o} , clay polymerization rate (of RNA) k_{clay-p} , and clay polymerization fidelity p . Therefore the full set of variables is given by $\theta = (n, k_{ss}, k_{ds}, k, k_{clay-o}, k_{clay-p}, p)$. The value of k and k_{clay-p} are set such that the replicative mass of each is initialized to 10. This means that RNA and clay polymerization have the same reaction mass at the beginning of the simulation. We set $n = 3$, $k = -\log(0.01)/n$, $p = 0.9$, and $k_{ss} = k_{ds} = 1$. This is a high hitting regime, i.e. $\mathbb{P}(\tau_v(\theta) < \infty) \sim 1$. In Figure 4 we plot measures of a typical realization of X_t on (H_i) and additionally the probability of reactions over time. High-fidelity replicators ended the simulation with 25% concentration (Figure 4a) and RNA polymerase output $\sim 62\%$ (Figure 4c), indicating the concentration of high-fidelity replicators will continue to increase. All species exhibit superexponential growth (Figure 4b). Clay polymerization decreases in contribution over time, whereas RNA polymerization increases substantially over time, and RNA double-strand reactions are small and stable (Figure 4d).

[Figure 4]

3.3 Hitting times: functional and survival analysis

We study various models in order of increasing complexity. We examine the hitting time surface $\tau_v(\theta)$ in $\theta \in \Theta$, including probability of hitting $\mathbb{P}(\tau_v(\theta) < \infty)$. We begin with the core model with no decay or clay.

3.3.1 Core model, $\tau_v(\theta)$ for $v = 0.1$ with $\theta = (n, k)$ and “tent” functions

We are interested in the structure of the hitting time $\tau_v(\theta)$ of (19) as a function of θ . We use the Weibull-Cox proportional hazard’s model of Table 1 for the hitting time τ_v for $v = 0.1$. Let $\theta = (n, k)$ with genome dimension $n \in \{3, 4\}$ and fitness and similarity parameters $k = l = -\log(i)/n$ for $i \in \{0.1, 0.05, 0.01, 0.005, 0.001\}$. Set $m = -\log(0.25)/n$ for fidelity probability parameter. For each value of n , take random X_0 with $I = |X_0| = 10$ and random singleton $R = \{\{x\}\}$ and fix these for the k . We fix the parameters $k_{ss} = k_{ds} = 1$ and set k_{rep} such that $k_{rep}(0) = 10$ and use the “tent” function for fitness, similarity, and fidelity. We take 10 realizations of $\tau_v(\theta)$ for each $\theta \in \Theta$ and allocate 5 000 reactions. This gives 100 independent hitting times and up to 500 000 reactions. *The times are comparable because the system is initialized to the same replication mass.*

For the simulations, 66 hitting times are finite. The coefficients positively contribute to hitting, where $\gamma_n \approx 0.97$ and $\gamma_k \approx 13.29$, both with p -values less than 0.005. Therefore hittings are strongly positively influenced by the parameters of the fitness and similarity functions and less so by the dimension. Plots of the coefficients and survival and cumulative hazard curves are given in Figure 5. *High survival is found for k large and low survival for k small. Cumulative hazard is highest for k small.*

[Figure 5]

We estimate HDMR of the classifier (whether or not hitting time is finite) using all 100 samples. The results are shown in Figure 6. The HDMR explains roughly 80% of variance. $S_k \approx 0.69$ and $S_n \approx 0.06$, so *hitting probability is strongly influenced by k and less so by n* . The component functions f_k and f_n are strictly decreasing, where larger k results in decreasing hitting probability. These results are consistent with the survival analysis.

[Figure 6]

We estimate HDMR of the regressor (hitting time) using the 66 simulations with finite hitting time. The results are shown in Figure 7. The HDMR explains roughly 60% of variance. $S_n \approx 0.57$ and $S_k \approx 0.04$, so *genome dimension dominates the hitting time*. Both f_n and f_k are increasing. The HDMR results reveal that conditioning on hitting reverses the roles of n and k .

[Figure 7]

3.3.2 Clay and decay model, $\tau_v(\theta)$ for $v = 0.1$ with $\theta = (n, k, k_\emptyset, k_{clay-p}, p)$ and “tent” functions

We expand the model to include clay and decay. We take

$$\begin{aligned}\theta &= (n, k, k_\emptyset, f_{clay}, p_{clay}) \in \Theta \\ \Theta &= \{3, 4\} \times \{-\log(i)/n : i = 0.1, 0.05, 0.01, 0.005, 0.001\} \times (0, 1) \times (0, 1) \times (0, 1)\end{aligned}$$

with $k_{ss} = k_{ds} = k_{clay-o} = 1$. For each $\theta \in \Theta$, (i) we choose random $R = \{x\}$, $x \in E$ and choose random X_0 such that $I = |X_0| = 10$ and $X_0 \cap R = \emptyset$, that is, the initial population does not reside on the high-fidelity manifold; (ii) we initialize the replicative mass of the system such that $k_{rep}(0) = (1 - f_{clay})20$ and $k_{clay-p}(0) = f_{clay}20$; (iii) we sample $\tau_v(\theta)$ for $v = 0.10$ $M = 10$ times, each censored by 5 000 reactions, giving $\mathcal{T}(\theta)$ of (20). We attain $\mathcal{D} = \{(\theta_i, \mathcal{T}(\theta_i)) : i = 1, \dots, 240\}$. This gives a total of 2 400 simulations.

For the simulations, 1 546 hitting times are finite. The results of fitting the Weibull-Cox model are shown below in Table 4 and Figure 8. The curvature parameter k again significantly dominates with a large positive value. All the remaining parameters have values less than one. n again is a relatively small positive contributor. The clay fraction f_{clay} is small and positive and p is not significant.

[Table 4]

[Figure 8]

We estimate HDMR of the classifier (whether or not hitting time is finite) using all 2 400 samples. Component functions and sensitivity indices are shown below in Figure 9. First-order HDMR captures 74% of explained variance, and second-order captures 4%. *Curvature dominates hitting probability with large sensitivity index* $\mathbb{S}_k \approx 67\%$. f_k is a decreasing function, where small values increase and large values decrease hitting probability. Genome dimension n has sensitivity index $\mathbb{S}_n \approx 2\%$, and f_n is decreasing, where high dimension decreases the probability of hitting. Clay parameter $\mathbb{S}_{f_{clay}} \approx 2\%$, and $f_{f_{clay}}$ is decreasing, where low-to-medium clay fractions increase and high-clay fractions decrease probability of hitting. The HDMR results are consistent with the Weibull-Cox model.

[Figure 9]

We estimate HDMR of the regressor (hitting time) using the 1 546 simulations with finite hitting time. Component functions and sensitivity indices are shown below in Figure 10. First-order HDMR captures 33% of explained variance, and second-order captures 7%. *In stark contrast to the contributions to the classifier, the parameters k and n are insignificant to hitting time. Instead, the largest sensitivity index is $\mathbb{S}_{f_{clay}} \approx 20\%$. $f_{f_{clay}}$ is an increasing function, where small f_{clay} decreases and large f_{clay} increases the hitting*

time. This suggests that high clay-fractions representing first-order reactions increase the hitting time, as clay polymerization has less replicative mass than RNA polymerization, i.e. things go faster with RNA polymerization. The second largest is decay $S_{k\emptyset} \approx 11\%$. Decay is an increasing function, with sharp increase in hitting times nearby one, i.e. things go slower with large decay resulting in increased hitting time.

[Figure 10]

3.4 Reaction cardinality

In this section we study, instead of the hitting time $\tau_v(\theta)$, the hitting reaction number $\varpi_v(\theta)$ for $v = 0.1$. We study the core model with $\theta = (n, k)$. We uniformly sample $n \in \{3, 4, 5\}$ and $k \in \{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$ and form data $\mathfrak{D} = \{(\theta_i, \varpi(\theta_i)) : i = 1, \dots, 1200\}$. We set $k_{ss} = k_{ds} = 1$ and initialize k_{rep} such that the initial replicative mass is 10. For each $\theta \in \Theta$, we randomly choose X_0 with $I = |X_0| = 10$ and singleton R such that $X_0 \cap R = \emptyset$.

$\mathfrak{D}$ contains 781 hitting events. We denote these $\mathfrak{D}^*$. We form a first-order HDMR on $\varpi(\theta)$ using $\mathfrak{D}^*$. The HDMR truth-plot, component functions, and sensitivity indices are shown below in Figure 11. First-order HDMR captures approximately 31% of variance. Both component functions f_n and f_k have similar sizes, with f_n somewhat larger than f_k . f_n is essentially an increasing linear function of n , and f_k is a generally an increasing function. *These results suggest that genome dimension and curvature influence the hitting reaction.* Larger genomes and flatter curvature increase the hitting reaction.

[Figure 11]

3.5 Compartmentalization

Compartmentalization has a direct effect on the calculation of the reaction rates, specifically replication, by computing only a subset of the reactions in $\mathcal{X}_1^2$. Put

$$X_t(A) = \{x \in X_t : l(x) \in A\}.$$

For micelle region $A \in \mathfrak{M}$ we have that

$$\begin{aligned} \check{k}_{rep}(t, A) &= \sum_{(x,y) \in \mathcal{X}_1^2} k_{rep}(x, y) M_t(\{x\} \times A) (M_t(\{y\} \times A) - \mathbb{I}(x = y)) \quad \text{for } t \in \mathbb{R}_+, \quad A \subset [-T, T] \\ &= \sum_{(x,y) \in X_t^2(A)} k_{rep}(x, y) M_t(\{x\} \times A) (M_t(\{y\} \times A) - \mathbb{I}(x = y)) \end{aligned}$$

and total replicative mass

$$\check{k}_{rep}(t) = \sum_{A \in \mathfrak{M}} \check{k}_{rep}(t, A) \quad \text{for } t \in \mathbb{R}_+$$

As $\mathfrak{M}$ increases in size over time, the number of partitions grows, and

$$\sum_{A \in \mathfrak{M}} |X_t^2(A)| \ll |\mathcal{X}_1^2|.$$

Therefore the replicative mass will be reduced with $\mathfrak{M}$, and the system evolves less quickly. *This suggests that compartmentalization should occur after replication. If it co-occurs with the RNA search, then it slows the system down.*

4 Conclusions

Origins of life is a fascinating problem. The wonderful complexity of extant life follows from origins. The distribution of life in the universe is tied to origins.

In this article, we have attempted to “peek” into the problem by concentrating on the RNA world hypothesis, studying hitting times of high-fidelity replicators. We develop fitness, similarity, and fidelity functions as landscapes for a mathematical model of replicating RNA molecules at the genome level and observe hitting times through simulation studies. We draw attention to the distinction between the probability of hitting $\mathbb{P}(\tau(\theta) < \infty)$ and the hitting time $\tau(\theta) < \infty$.

In terms of mathematical set-up, we interpret the reactions as measure-kernel-functions. Each reaction is identified to and fully encoded by a probability transition kernel. This applies as well to other ‘privileged’ functions, with corresponding transition kernels, although these in particular remain abstract and are generally outside the intended scope of this article. The reactions take place in some domain, whereby all molecules may interact. We note that, in reality, molecules have limited diffusion, and this effectively breaks the reaction domain into independent subdomains above some length scale, i.e. molecules are more likely to react with their neighbors. Therefore we assume our reaction volume is sufficiently small such that all molecules may participate in the reactions.

We find that *non-linear landscapes* are *necessary* for hitting: linear landscapes are insufficient. For non-linear landscapes, we find that the *probability of hitting* is dominated by *curvature* and that *hitting times* are dominated by *genome dimension*. These results suggest that the landscapes in nature are non-linear with high curvature, and that the hitting time for high-fidelity replicators is an increasing function of genome dimension. When clay and decay are added to the model, hitting probability is again dominated by curvature, and clay and decay are relatively insignificant. However clay and decay significantly contribute to the hitting times, in both cases increasing hitting times.

In terms of the relation between replication and compartmentalization and metabolism, we find that compartmentalization slows system dynamics and should occur after the RNA search is concluded. Therefore compartmentalization is identified to a single transition kernel Q_c , tied to genomic adaptation. It is possible to argue that because the system takes place inside some domain, i.e. a compartment, compartmentalization is a necessary

condition. Metabolism is thought to be identified to production of precursors to RNA synthesis, leading to replication identification, followed by genomic adaptation to metabolism. Metabolism is thus defined through two transition kernels, Q_m and Q_e . The independence of these transition kernels can be called into question, and it is quite possible that general transition kernels on the full product spaces may be necessary to satisfactorily explain origins. That is, all three ‘privileged’ functions may have co-evolved.

If we let $f \in \mathcal{E}_+$ be a fitness function, then the fitness value $J(\mu_t) = \mu_t f$ is the expected value of the fitness function with respect to the probability measure $\mu_t = \nu_t Q$. In OptiEvo theory, $J(\mu_t)$ is studied as a function of the population X_t on $(E, \mathcal{E})$ (Feng et al., 2012). OptiEvo assumes that the set of all probability measures $\{\mu_t\}$ is convex and that X_t has sufficient flexibility such that $J(\mu_t)$ may be explored around μ_t . Then OptiEvo predicts that $J(\mu_t)$ has global maxima on $(E, \mathcal{E})$ and that these form a connected level-set of genomes with the same fitness value. Both predictions hold for our model. The first prediction is satisfied by zero distance in fitness and similarity functions for high-fidelity genomes. The second prediction is implied by the high-fidelity set being a singleton or a product-space construction. The main point of contention is whether X_t has sufficient flexibility in exploring $J(\mu_t)$ around μ_t . This has direct bearing on the structure of Q : if X_t is inflexible, then Q is constrained to certain subspaces of $(E, \mathcal{E})$, i.e. not all transitions are possible.

For future work, the model could be extended to the space of genomes of lengths up to n , $(E^*, \mathcal{E}^*)$ or even the space of genomes of all lengths, where distance and similarity functions would utilize a more general string distance metric, e.g., Levenshtein distance. We note that the size of $(E^*, \mathcal{E}^*)$ is not much larger than $(E, \mathcal{E})$. The aforementioned transition kernels Q_c , Q_m , and Q_e could be studied. More general models for polymerization Q based on sequence-specific schemata could be developed, using the structure of the Poisson-binomial distribution. It would be interesting to study lipid-RNA and metabolism-RNA interactions and equip the system with the ability to append nucleotides to their genomes to form genes, such as storing useful information for the replication channel. For example, genomes could become adapted relative to schemata that code for reproductively useful functions, e.g., a “primordial shield” hypothesis of lipids that interact with specific genomes which confers protection against decay through shielding.

5 Figures

Figure 1: Measures of X_t until τ_v for $v = 0.25$ with $n = 3$ and $k = -\log(0.01)/n$ and $I = |X_0| = 10$ and $R = \{\{x\}\}$ and “tent” fitness and similarity functions

Figure 2: Measures of X_t until τ_v for $v = 0.25$ with $n = 3$ and $k = -\log(0.1)/n$ and $I = |X_0| = 10$ and $R = \{\{x\}\}$ and “tent” fitness and similarity functions

Figure 3: Measures of X_t until τ_v for $v = 0.25$ with $n = 3$ and $k = -\log(0.01)/n$ and $I = |X_0| = 10$ and $R = \{\{x\}\}$ and linear fitness and similarity functions

Figure 4: Measures of X_t until $\tau_v(\theta)$ for $v = 0.25$ with $n = 3$ and $k = -\log(0.01)/n$ and $I = |X_0| = 10$ and $R = \{\{x\}\}$ and $k_{ss} = k_{ds} = 1$ and $p = 0.9$ and k_{rep}, k_{clay-p} chosen such that the replicative mass of each is 10

(a) Coefficients

(b) Survival curves

(c) Cumulative hazard

Figure 5: Survival analysis of hitting time τ_v for $v = 0.1$ for core model

(a) Hexagonal-bin truth plot

(b) $f_n(n)$

(c) $f_k(k)$

θ	S_θ
n	0.06
k	0.69
Σ	0.75

(d) First-order sensitivity indices

Figure 6: First-order HDMR analysis of $\mathbb{P}(\tau_v(\theta) < \infty)$ for core model

(a) Hexagonal-bin truth plot

(b) $f_n(n)$

(c) $f_k(k)$

θ	S_θ
n	0.57
k	0.04
Σ	0.61

(d) First-order sensitivity indices

Figure 7: First-order HDMR analysis of $\tau_v(\theta) < \infty$ for core model

(a) Coefficients

(b) Survival curves

(c) Cumulative hazard

Figure 8: Survival analysis of hitting time τ_v for $v = 0.1$ for expanded model (clay and decay)

(a) Hexagonal-bin truth plot

(b) $f_n(n)$

(c) $f_k(k)$

(d) $f_{f_{clay}}(f_{clay})$

θ	$\mathbb{S}_\theta$
n	0.0213
k	0.6732
$k_\emptyset$	0.0120
p	0.0114
f_{clay}	0.0219
$\sum$	0.7399

(e) First-order sensitivity indices

Figure 9: First-order HDMR analysis of $\mathbb{P}(\tau_v(\theta) < \infty)$ for expanded model (clay and decay)

(a) Hexagonal-bin truth plot

(b) $f_n(n)$

(c) $f_k(k)$

(d) $f_{f_{clay}}(f_{clay})$

(e) $f_{k_{\emptyset}}(k_{\emptyset})$

θ	$\mathbb{S}_{\theta}$
n	0.0013
k	0.0030
$k_{\emptyset}$	0.1129
p	0.0152
f_{clay}	0.2014
Σ	0.3339

(f) First-order sensitivity indices

Figure 10: First-order HDMR analysis of $\tau_v(\theta) < \infty$ for expanded model (clay and decay)

(a) Hexagonal-bin truth plot

(b) $f_n(n)$

(c) $f_k(k)$

θ	$\mathbb{S}_\theta$
n	0.1619
k	0.1464
Σ	0.3083

(d) First-order sensitivity indices

Figure 11: First-order HDMR analysis of $\varpi_v(\theta) < \infty$ for core model and $v = 0.1$

Figure 12: First-order HDMR analysis of $f(n, r, m, k, l)$

6 Tables

Table 1: Weibull reliability model

Name	Quantity	Baseline	Quantity	Proportional Hazards
Failure density	$f(t \alpha, \beta)$	$\frac{\alpha}{t}(\frac{t}{\beta})^{\alpha}e^{-(\frac{t}{\beta})^{\alpha}}$	$f(t, \theta \alpha, \beta, \gamma)$	$\frac{\alpha}{t}(\frac{t}{\beta})^{\alpha}e^{\gamma \cdot \theta}e^{-(\frac{t}{\beta})^{\alpha}e^{\gamma \cdot \theta}}$
Failure distribution	$F(t \alpha, \beta)$	$1 - e^{-(\frac{t}{\beta})^{\alpha}}$	$F(t, \theta \alpha, \beta, \gamma)$	$1 - e^{-(\frac{t}{\beta})^{\alpha}e^{\gamma \cdot \theta}}$
Reliability distribution	$R(t \alpha, \beta)$	$e^{-(\frac{t}{\beta})^{\alpha}}$	$R(t, \theta \alpha, \beta, \gamma)$	$e^{-(\frac{t}{\beta})^{\alpha}e^{\gamma \cdot \theta}}$
Cumulative hazard	$H(t \alpha, \beta)$	$(\frac{t}{\beta})^{\alpha}$	$H(t, \theta \alpha, \beta, \gamma)$	$(\frac{t}{\beta})^{\alpha}e^{\gamma \cdot \theta}$
Hazard rate	$h(t \alpha, \beta)$	$\frac{\alpha}{t}(\frac{t}{\beta})^{\alpha}$	$h(t, \theta \alpha, \beta, \gamma)$	$\frac{\alpha}{t}(\frac{t}{\beta})^{\alpha}e^{\gamma \cdot \theta}$

Table 2: Reactions and orders

Reaction	Order
RNA double strand formation	2
RNA double strand dissociation	1
RNA polymerization	2
RNA decay	1
Clay polymerization	1
Clay oligomerization	0
Compartmentalization	1
Metabolism to replication	1
Metabolism & replication to metabolism	2

Table 3: Model parameters

θ	Name	Domain
n	genome dimension	$\mathbb{N}_{>0}$
q	R probability of subset size 4 for position	$[0, 1]$
k	RNA fitness parameter	$\mathbb{R}_+$
l	RNA similarity parameter	$\mathbb{R}_+$
m	RNA fidelity parameter	$\mathbb{R}_+$
p	clay fidelity probability	$(0, 1]$
k_{ss}	double-strand dissociation rate	$\mathbb{R}_+$
k_{ds}	double-strand formation rate	$\mathbb{R}_+$
k_{rep}	RNA replication rate	$\mathbb{R}_+$
$k_{\emptyset}$	RNA decay rate	$\mathbb{R}_+$
k_{clay-o}	clay RNA oligomerization rate	$\mathbb{R}_+$
k_{clay-p}	clay RNA polymerization rate	$\mathbb{R}_+$

Table 4: Weibull-Cox model parameters for hitting times of clay and decay model

θ	Name	Coefficient γ_{θ}	p -value
n	genome dimension	0.54	< 0.005
$k = l$	RNA fitness parameter	27.87	< 0.005
p	clay fidelity probability	-0.08	0.35
$k_{\emptyset}$	RNA decay rate	0.80	< 0.005
f_{clay}	fraction clay RNA polymerization rate	0.96	< 0.005

A Discrete probability space

We define some concepts related to the space $(E, \mathcal{E})$. The discrete probability measure ν on $(E, \mathcal{E})$ is defined by

$$\nu(A) = \nu \mathbb{I}_A = \sum_{x \in E} \nu\{x\} \mathbb{I}_A(x) \quad \text{for } A \in \mathcal{E}$$

where $\nu\{x\}$ is the probability mass at the point $x \in E$. For the collection of positive $\mathcal{E}$ -measurable functions $\mathcal{E}_+$ we have

$$\nu f = \sum_{x \in E} \nu\{x\} f(x) \quad \text{for } f \in \mathcal{E}_+.$$

B Other fitness functions

Another fitness function can be defined using polynomials, such as lines, quadratics, etc, in terms of $k \in \mathbb{N}_{>0}$

$$f_k(x, R) = \left(1 - \frac{H(x, R)}{n+1}\right)^k \in (0, 1] \quad \text{for } x \in E$$

or

$$f_k(x, R) = 1 - \left(\frac{H(x, R)}{n+1}\right)^k \in (0, 1] \quad \text{for } x \in E$$

Yet another surface is using a sigmoid function. Put

$$E(x) = \frac{1}{1 + \exp[-x]} \quad \text{for } x \in \mathbb{R}$$

We have fitness for $k \in (0, \infty)$

$$f_k(x, R) = \frac{1 - E\left(\frac{H(x, R) - \frac{n}{2}}{k}\right)}{1 - E\left(-\frac{n}{2k}\right)} \in (0, 1] \quad \text{for } x \in E$$

C Measure-kernel-function

We recall a few facts about transition kernel Q . Q defines a function

$$Qf(x) = \int_Y Q(x, dy) f(y) \quad \text{for } x \in E$$

that is in $\mathcal{X}_+$ for every function $f \in \mathcal{Y}_+$. For every probability measure ν_t on $(E, \mathcal{E})$ the quantity $\mu_t = \nu_t Q$ defines a probability measure on $(Y, \mathcal{Y})$ as

$$\mu_t(A) = \int_X \nu_t(dx) Q(x, A) \quad \text{for } A \in \mathcal{Y}.$$

For every probability measure ν_t on $(X, \mathcal{X})$ and function $f \in \mathcal{Y}_+$ we have that

$$\mu_t f = (\nu_t Q) f = \int_X \nu_t(dx) \int_Y Q(x, dy) f(y).$$

Here the spaces are discrete, i.e.

$$\nu_t(A) = \nu_t \mathbb{I}_A = \sum_{x \in X} \nu_t\{x\} \mathbb{I}_A(x) \quad \text{for } A \in \mathcal{X}$$

where $\nu_t\{x\}$ is the probability mass at the point $x \in X$.

D Approximate reaction rates

One approach to reducing the computational complexity of $\check{k}(t)$ is to approximate the sums using Monte Carlo. Define random variables $\mathfrak{X}_1 \sim \text{Uniform}(\mathcal{X}_1)$ and $\mathfrak{X}_2 \sim \text{Uniform}(\mathcal{X}_2)$. Let $\{\mathfrak{X}_{1i}\}$ and $\{\mathfrak{X}_{2i}\}$ be independencies of such random variables. Given N random samples of $\mathfrak{X}_1$, the first reaction rate becomes

$$\check{k}_{ds}(t|N) = \frac{|\mathcal{X}_1|}{2N} \sum_{i=1}^N k_{ds} N_t(\mathfrak{X}_{1i}) N_t(\mathfrak{X}_{1i}^c)$$

whose expected value is approximated using M realizations,

$$\mathbb{E} \check{k}_{ds}(t|M, N) \simeq \frac{|\mathcal{X}_1|}{2MN} \sum_{j=1}^M \sum_{i=1}^N k_{ds} N_t(\mathfrak{X}_{1ij}) N_t(\mathfrak{X}_{1ij}^c)$$

requiring a total of MN evaluations. In a similar manner, the second reaction rate is

$$\mathbb{E} \check{k}_{ss}(t|M, N) \simeq \frac{|\mathcal{X}_2|}{MN} \sum_{j=1}^M \sum_{i=1}^N k_{ss} N_t(\mathfrak{X}_{2ij} \cup \mathfrak{X}_{2ij}^c)$$

and putting $\mathfrak{X}_1^* \sim \text{Uniform}(\mathcal{X}_1)$ we have the third reaction

$$\mathbb{E} \check{k}_{rep}(t|M, N) \simeq \frac{|\mathcal{X}_1|^2}{MN} \sum_{j=1}^M \sum_{i=1}^N k_{rep}(\mathfrak{X}_{1ij}, \mathfrak{X}_{1ij}^*) N_t(\mathfrak{X}_{1ij}) (N_t(\mathfrak{X}_{1ij}^*) - \mathbb{I}(\mathfrak{X}_{1ij} = \mathfrak{X}_{1ij}^*)) \quad (34)$$

We refer to SSA simulation with Monte Carlo approximate reaction rates as Monte Carlo Approximate SSA, or MCASSA.

MCASSA performance analysis At any time point in time, the reaction rates are determined from the population X_t . We focus on the replication rate. We study the Monte Carlo estimator performance by varying the population X_t of single-stranded genomes of size m , genome dimension n , q (which is used to construct R), k (fitness degree), and l (similarity degree). We take $\theta = (n, q, m, k, l) \in \Theta = \{1, \dots, 10^2\} \times (0, 1) \times \{1, \dots, 10^3\} \times \{1, \dots, 6\} \times \{1, \dots, 6\}$ with uniform $\nu = \nu_n \nu_q \nu_m \nu_k \nu_l$. We generate input data $\{\theta_i\}$. For each $\theta_i \in \Theta$, we attain R_i . Then, given θ_i and R_i , we estimate (34) 100 times using Monte Carlo with $MN = 10^3$ and compute the mean to standard deviation (signal-to-noise ratio) $f(\theta_i) = \frac{\mu}{\sigma}(\theta_i)$ of the samples. f encodes the performance or efficiency of the reaction rate estimator. This procedure is repeated 10^3 times to generate input-output data $\mathcal{D} = \{(\theta_i, f(\theta_i)) : i = 1, \dots, 10^3\}$. From $\mathcal{D}$ we estimate first-order HDMR of f

$$f(n, q, m, k, l) = f_0 + f_n(n) + f_q(q) + f_m(m) + f_k(k) + f_l(l)$$

using degree 5 polynomials in each variable and then compute the sensitivity indices $\mathbb{S}_n$, $\mathbb{S}_q$, $\mathbb{S}_m$, $\mathbb{S}_k$, and $\mathbb{S}_l$. Truth plots and first-order component functions and sensitivity indices are shown below in Figure 12. First-order HDMR captures $\sim 74\%$ of variance. The largest contributor is similarity degree l with $\sim 48\%$ of variance, where increasing l decreases the signal-to-noise ratio. The next largest contributor is genome dimension n with $\sim 13\%$ of variance, where increasing n increases the signal-to-noise ratio. The next largest contributor is the fraction q with $\sim 9\%$ of variance, where increasing q increases the signal-to-noise ratio. The next largest contributor is fitness degree k with $\sim 4\%$ of variance, where increasing k decreases the signal-to-noise ratio. The smallest contributor by far is population size m with $\sim 0.04\%$ of explained variance. This shows that the estimator performance is generally independent of the population size.

[Figure]

E High dimensional model representation

Suppose we have a real-valued square-integrable function $f(x) \in L^2(E, \mathcal{E}, \nu)$ with $E = \mathbb{R}^n$, $\mathcal{E} = \mathcal{B}_{\mathbb{R}^n}$, $\nu = \prod_i \nu_i$, and $x = (x_1, \dots, x_n) \in E$. The $\{\nu_i\}$ may be diffuse (continuous) and/or discrete. Put $B = \{1, \dots, n\}$. We would like to decompose f into orthogonal function subspaces $\{f_u : u \subseteq B\}$ (projections) in such a way that each projection on an input subspace f_u maximizes variance and across subspaces retrieves total variance, i.e. $f = \sum_{u \subseteq B} f_u$ and $\mathbb{V}ar f = \sum_{u \subseteq B} \mathbb{V}ar f_u$. The solution to this problem in the retrieval of $\{f_u : u \subseteq B\}$ is known as *high dimensional model representation* (HDMR) or functional ANOVA expansion and for f is written as

$$f(x_1, \dots, x_n) = f_0 + \sum_i f_i(x_i) + \sum_{i < j} f_{ij}(x_i, x_j) + \dots + f_{1\dots n}(x_1, \dots, x_n)$$

where the $\{f_u : u \subseteq B\}$ are called *component functions*. For independent inputs, the component functions are mutually orthogonal and, aside from the constant component function $f_0 = \mathbb{E}f$ (order zero), have zero mean $\mathbb{E}f_u = 0$ for all non-empty $(2^n - 1)$ subspaces, where

$$\mathbb{V}ar f_u = \int_E f_u^2(x_u) \nu(dx) < \infty \quad \text{for } u \subseteq B$$

and

$$\mathbb{V}ar f = \int_E (f(x) - f_0)^2 \nu(dx) = \sum_{u \subseteq B} \mathbb{V}ar f_u < \infty.$$

A key principle of HDMR is that the expansion for most f may be truncated at low order $T \ll n$ in a T -order HDMR,

$$f(x) \simeq f^T(x) = \sum_{u \subseteq B: |u| \leq T} f_u(x_u) \quad \text{for } T \ll n.$$

HDMR is often used in *global sensitivity analysis* to assess input-output correlations at various orders, where the variances are normalized to define *sensitivity indices*

$$\mathbb{S}_u = \frac{\mathbb{V}ar f_u}{\mathbb{V}ar f} \quad \text{for } u \subseteq B.$$

When the inputs are correlated $\nu \neq \prod_i \nu_i$, then the component functions may still be uniquely recovered under hierarchical orthogonality, the variance decomposes

$$\mathbb{V}ar f = \sum_{u, v \subseteq B} \mathbb{C}ov(f_u, f_v),$$

where

$$\mathbb{C}ov(f_u, f_v) = \int_E f_u(x_u) f_v(x_v) \nu(dx) \quad \text{for } u, v \subseteq B$$

and the sensitivity indices generalize to *structural* and *correlative* sensitivity indices (Li and Rabitz, 2012), defined respectively as

$$\mathbb{S}_u^a = \frac{\mathbb{V}ar f_u}{\mathbb{V}ar f} \quad \text{for } u \subseteq B$$

and

$$\mathbb{S}_u^b = \frac{\sum_{v \subseteq B: u \neq v} \mathbb{C}ov(f_u, f_v)}{\mathbb{V}ar f} \quad \text{for } u \subseteq B$$

with *total* sensitivity index

$$\mathbb{S}_u = \mathbb{S}_u^a + \mathbb{S}_u^b \quad \text{for } u \subseteq B$$

where $\sum_{u \subseteq B} \mathbb{S}_u = 1$. We use the total sensitivity index as a measure of variable importance and the component functions as profiles of output dependence on the input subspaces.

F Reliability definitions

Given a failure distribution f and reliability (survival) distribution R , we give some relations: the cumulative failure distribution is defined as

$$F(t|\vartheta) = \int_0^t f(s|\vartheta)ds$$

where

$$R(t|\vartheta) + F(t|\vartheta) = 1,$$

the hazard rate $h(t|\vartheta)$ is defined as

$$h(t|\vartheta) = \frac{f(t|\vartheta)}{R(t|\vartheta)},$$

the cumulative hazard is defined as

$$H(t|\vartheta) = \int_0^t h(s|\vartheta)ds,$$

and we have reliability expressed in terms of the cumulative hazard

$$R(t|\vartheta) = e^{-H(t|\vartheta)}.$$

Another useful quantity is the mean residual life

$$\mu(t|\vartheta) = \frac{\int_t^\infty R(s|\vartheta)ds}{R(t|\vartheta)}.$$

G Declarations

Funding The authors did not receive support from any organization for the submitted work.

Conflicts of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Author contributions All authors contributed to the study conception and design. Material preparation, data generation and analysis were performed by Caleb Bastian. The first draft of the manuscript was written by Caleb Bastian and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval Not applicable

Consent to participate Not applicable

Consent for publication Not applicable

References

- Attwater, J., Wochner, A., and Holliger, P. (2013). In-ice evolution of rna polymerase ribozyme activity. *Nature chemistry*, 5(12):1011–1018.
- Bornholt, J., Lopez, R., Carmean, D. M., Ceze, L., Seelig, G., and Strauss, K. (2016). A dna-based archival storage system. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS ’16, pages 637–649, New York, NY, USA. Association for Computing Machinery.
- Cairns-Smith, A. G. (1987). *Genetic Takeover and the Mineral Origins of Life*. Cambridge University Press.
- Cech, T. R. (2000). The ribosome is a ribozyme. *Science*, 289(5481):878.
- Coveney, P. V., Swadling, J. B., Wattis, J. A. D., and Greenwell, H. C. (2012). Theory, modelling and simulation in origins of life studies. *Chemical Society Reviews*, 41(16):5430–5446.
- Davidson-Pilon, C. (2019). lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317.
- Diener, T. O. (1971). Potato spindle tuber ”virus”. iv. a replicating, low molecular weight rna. *Virology*, 45(2):411–428.
- Dyson, F. (1999). *Origins of Life*. Cambridge University Press, Cambridge.
- Ehrenfreund, P., Rasmussen, S., Cleaves, J., and Chen, L. (2006). Experimentally tracing the key steps in the origin of life: The aromatic world. *Astrobiology*, 6(3):490–520.
- Eigen, M. and Schuster, P. (1977). A principle of natural self-organization. *Naturwissenschaften*, 64(11):541–565.
- Ertem, G. and Ferris, J. P. (1997). Template-directed synthesis using the heterogeneous templates produced by montmorillonite catalysis. a possible bridge between the prebiotic and rna worlds. *Journal of the American Chemical Society*, 119(31):7197–7201.
- Feng, X., Pechen, A., Jha, A., Wu, R., and Rabitz, H. (2012). Global optimality of fitness landscapes in evolution. *Chemical Science*, 3(3):900–906.

- Ferris, J. P. (2006). Montmorillonite-catalysed formation of rna oligomers: the possible role of catalysis in the origins of life. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 361(1474):1777–1786.
- Gilbert, W. (1986). Origin of life: The rna world. *Nature*, 319(6055):618–618.
- Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.
- Hays, L. (2015). Nasa astrobiology strategy. Technical report, National Aeronautics and Space Administration.
- Hordijk, W. and Steel, M. (2004). Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *Journal of Theoretical Biology*, 227(4):451–461.
- Joyce, G. F. (2002). The antiquity of rna-based evolution. *Nature*, 418(6894):214–221.
- Kitadai, N. and Maruyama, S. (2018). Origins of building blocks of life: A review. *Geo-science Frontiers*, 9(4):1117–1153.
- Kunin, V. (2000). A system of two polymerases –a model for the origin of life. *Origins of life and evolution of the biosphere*, 30(5):459–466.
- KVENVOLDEN, K., LAWLESS, J., PERING, K., PETERSON, E., FLORES, J., PON-NAMPERUMA, C., KAPLAN, I. R., and MOORE, C. (1970). Evidence for extraterrestrial amino-acids and hydrocarbons in the murchison meteorite. *Nature*, 228(5275):923–926.
- Lanier, K. A. and Williams, L. D. (2017). The origin of life: Models and data. *Journal of Molecular Evolution*, 84(2).
- Li, G. and Rabitz, H. (2012). General formulation of hdmr component functions with independent and correlated variables. *Journal of Mathematical Chemistry*, 50(1):99–130.
- Martin, W., Baross, J., Kelley, D., and Russell, M. J. (2008). Hydrothermal vents and the origin of life. *Nature Reviews Microbiology*, 6(11):805–814.
- Maury, C. P. J. (2015). Origin of life. primordial genetics: Information transfer in a pre-rna world based on self-replicating beta-sheet amyloid conformers. *Journal of Theoretical Biology*, 382:292–297.
- Miller, S. L. and Urey, H. C. (1959). Organic compound synthesis on the primitive earth. *Science*, 130(3370):245.
- Orgel, L. (2000). A simpler nucleic acid. *Science*, 290(5495):1306.

- Orgel, L. E. (1968). Evolution of the genetic apparatus. *Journal of Molecular Biology*, 38(3):381–393.
- Sakuma, Y. and Imai, M. (2015). From vesicles to protocells: the roles of amphiphilic molecules. *Life (Basel, Switzerland)*, 5(1):651–675.
- Schmitt, A. O. and Herzel, H. (1997). Estimating the entropy of dna sequences. *Journal of Theoretical Biology*, 188(3):369–377.
- Segré, D., Ben-Eli, D., Deamer, D. W., and Lancet, D. (2001). The lipid world. *Origins of life and evolution of the biosphere*, 31(1):119–145.
- Szostak, J. W., Bartel, D. P., and Luisi, P. L. (2001). Synthesizing life. *Nature*, 409(6818):387–390.
- Vaidya, N., Manapat, M. L., Chen, I. A., Xulvi-Brunet, R., Hayden, E. J., and Lehman, N. (2012). Spontaneous network formation among cooperative rna replicators. *Nature*, 491(7422):72–77.
- Walker, S. I. (2017). Origins of life: a problem for physics, a key issues review. 80(9):092601.
- Wattis, J. A. D. and Coveney, P. V. (1999). The origin of the rna world: A kinetic model. *The Journal of Physical Chemistry B*, 103(21):4231–4250.
- Wright, S. (1984). *Evolution and the Genetics of Populations, Volume 1*. The University of Chicago Press.
- Wu, M. and Higgs, P. G. (2012). The origin of life is a spatially localized stochastic transition. *Biology Direct*, 7(1):42.